DOCUMENT RESUME

ED 369 820 TM 021 464

AUTHOR	Frederiksen, Norman
TITLE	The Influence of Minimum Competency Tests on Teaching and Learning.
INSTITUTION	Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE	Mar 94
NOTE	33p.
PUB TYPE	Information Analyses (070) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Tests; Age Differences; Educational History; Elementary Secondary Education; *Learning; Literature Reviews; *Mathematics Tests; *Minimum Competency Testing; National Surveys; Policy Formation; Standards; *Teaching Methods; Test Use
IDENTIFIERS	Competency Based Evaluation; High Stakes Tests; *National Assessment of Educational Progress; *Testing Effects

ABSTRACT

Past research on the effects of Minimum Competency Tests (MCT) on teaching and learning is reviewed, and the large database of the National Assessment of Educational Progress (NAEP) is used to shed more light on these effects. There seems to be little doubt that MCT, together with associated changes in instructional methods, has produced some substantial changes in student performance. The influence of state-mandated MCT on the quality of teaching and learning as reflected in the NAEP was investigated by comparing the performance of participants in the 1978 mathematics assessment with performance of participants in the 1986 assessment; the same set of items was used on both occasions. The 1978 assessment occurred before MCT were in general use. The better performance of students in 1986 was probably due to the efforts of teachers who made use of MCTs and high-stakes tests. The younger students in 1986 apparently profited more from the MCTs and high-stakes tests than the older students did. It seems reasonable to conclude that the use of MCTs can have desirable influences on performance of young students as measured by the NAEP. Nine tables, including some in an appendix titled "Some Broader NEP Methods and Interests," present study findings. (Contains 16 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

A POLICY ISSUE PERSPECTIVE

POLICY INFORMATION CENTER
Educational Testing Service

The Influence of Minimum Competency Tests on Teaching and Learning

by Norman Frederiksen

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RICHARD J. COLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Copyright © 1994 by Educational Testing Service. All rights reserved.

The Influence of Minimum Competency Tests on Teaching and Learning

By

Norman Frederiksen
Educational Testing Service

March 1994

ETS
Policy Information Center
Princeton, NJ 08541

Table of Contents

	Page
Preface	i
Acknowledgments	i
Introduction	1
Changes in Educational Priorities	1
Changes in Assessment Methods	1
The Decline in School Achievement	2
Holding Schools Accountable	3
Some Evidence Regarding the Influence of MCTs on Teaching and Learning	3
A Nationwide Study of the Influence of State-Mandated Testing Practices	6
NAEP Policies in Relation to this Study	7
The Analysis of the NAEP Data	10
Some Interpretations of Our Findings	12
Appendix: Some Broader NAEP Methods and Interests	15
Bibliography	25

ERRATUM

In Table 6 (page 17), Table 7 (page 20), Table 8 (page 22), and Table 9 (page 25) of the Appendix section, the following symbol ° should be replaced with an Ω (Omega) symbol.

Preface

In *The Influence of Minimum Competency Tests on Teaching and Learning*, Norman Frederiksen reviews past research on the effects of Minimum Competency Tests (MCTs) on teaching and learning, and uses the large database of the National Assessment of Educational Progress to shed more light on the matter. Using NAEP trend data, he is able to contrast achievement changes in states using MCTs and assigning high stakes to them, with changes in states not using them. Testing, whether traditional or "reformed," remains central to current education reform efforts, and the issues explored by Frederiksen are as important today as they were in the days of "back to basics" and minimal competency testing. He has also pioneered the use of the important but complex NAEP database to conduct policy-relevant secondary research.

Paul E. Barton
Director
Policy Information Center

Acknowledgments

Robert Mislevy commented on the research design, while Dick Lesh had the important task of reviewing the NAEP items used in the study, in order to categorize them as "routine" or "nonroutine." Margaret E. Goertz similarly helped by setting standards for selection of "high-stakes" and "low-stakes" states. John Barone found that prior to the 1986 testing it had not been possible to identify the states in which the students were tested; however, it was found by Barone that the states could be identified in the 1986 and the 1990 testing, which made it possible to use both; the zip codes had been the solution to the problem. Gene Johnson and Stephen Koffler were important advisors at the early meetings, while Nancy Robertson provided the computer assistance; Laura J. Jerry and John Ferris were the helpers in the last stages of the project. Gene Johnson reviewed the completed manuscript. Gwen Shrift did the editing, and Carla Cooper provided desktop publishing services.

Introduction

This article describes how minimum competency tests affect teaching and learning.

It is usually assumed that curriculum leads instruction, and instruction leads testing (Tyler, 1934). Such relationships can easily be reversed. The state-mandated use of minimum-competency tests (MCTs) has influenced many schools to "teach for the test"—even to put aside the curriculum and lesson plans in order to prepare students for the MCTs. The result is what Airasian (1988) called "measurement-driven instruction"—a condition in which greater efforts are given to teaching, whatever knowledge and skills are being assessed by a test.

Changes in Educational Priorities

Mass education in America at the beginning was primarily an elementary school system that stressed the routine abilities needed for "readin', 'ritin', and 'rithmetic." Secondary school systems were later added to the system to teach more advanced knowledge and skills, but they soon became diversified by adding vocational and general programs along with the traditional disciplines of the more elite schools. But even the academic courses became watered down; for example, written composition has "all but disappeared from the curriculum" (Resnick, 1987, p. 6). According to Resnick, "The effect of all this has been to reduce, and sometimes to drive out of existence, the high literacy goals that had been the focus of the academies and preparatory institutions" (p. 6).

When the states assumed responsibility for mass education, their departments of education began to formulate curricula for the schools. Some states merely listed the subject-matter areas to be covered (e.g., reading, writing, mathematics, science, civics, and American history), while others prescribed in more detail the knowledge and skills to be taught.

Teachers were relieved of some of the chore of grading examinations when standardized achievement tests began to appear in multiple-choice form, which made it possible to score a test by counting the dots that were visible through the holes in a scoring key.

Changes in Assessment Methods

The only tests were those prepared by teachers. Some were daily or weekly quizzes to keep the students on their toes, and some were final examinations that were used to help the teachers decide what grades to enter on students' report cards. The tests were almost always composed of questions or problems that required written responses.

Teachers were relieved of some of the chore of grading examinations when standardized achievement tests began

to appear in multiple-choice form, which made it possible to score a test by counting the dots that were visible through the holes in a scoring key. The first standardized achievement test to be used nationally was probably the Stanford Achievement Test, which appeared in 1923. When the IBM test-scoring machines of World War II vintage were replaced by more sophisticated electronic test-scoring machines (Lindquist, 1954), standardized multiple-choice achievement tests were off to the races. Such tests were intended primarily for use in school evaluation. At that time, they had relatively little influence on teaching and learning.

ESEA legislation was unintentionally responsible for a change in the influence of tests on teaching.

The Decline in School Achievement

After the launch of Sputnik by the USSR in 1957, there was much concern about the quality of science education in the United States. Congress responded by passing the National Defense Act to support improved education in math and science, and in 1965 the Elementary and Secondary Education Act (ESEA) was passed to support efforts to improve education generally. According to Popham (1983), the ESEA legislation was unintentionally responsible for a change in the influence of tests on teaching. ESEA required that the results of the studies it funded be evaluated, which usually involved a standardized test. Future funding was likely to depend on the scores derived from these tests. This provided motivation to "teach for the test." As Popham (1983) described it, "Big dollars were riding on the results of achievement tests . . . The days of penny-ante assessment were over" (p. 23).

In spite of ESEA and other efforts to improve school performance, scores on various achievement tests revealed that knowledge and skills were indeed declining during the 1970s

In spite of ESEA and other efforts to improve school performance, scores on various achievement tests revealed that knowledge and skills were indeed declining during the 1970s (Womer, 1981). A study by Rock, Ekstrom, Goertz, Hilton, and Pollack (1985) provides a good account of the decline in schools' achievement. The data, which were provided by the National Longitudinal Study and the High School and Beyond study (Donlon, Hilton, & Schrader, 1978; Hilton & Rhett, 1973) concluded that there were indeed significant drops in performance. In standard deviation units, the drops over a period of several years were .22, .21, and .40 for vocabulary, reading, and mathematics, respectively. Some likely causes were identified: (a) a larger number of students elected general or vocational programs, (b) fewer students took college preparatory courses, (c) the amount of homework done decreased, and (d) there was less emphasis on academics in the schools.

In 1983 the Secretary of Education appointed a National Commission on Excellence in Education to look into the problem; the committee's report decried the "rising tide of mediocrity." At about the same time, the National Task Force on Education for Economic Growth concluded that schools were in serious trouble and that declining performance undermined efforts to improve the nation's economic position. Many suggestions for improving our schools came from these and other sources: ideas such as increasing teachers' salaries, keeping schools open year-round, decreasing class size, training teachers better, and "restoring a decaying society." But few if any suggestions specified in detail how the process of instruction within the classroom might be improved.

. . .few if any suggestions specified in detail how the process of instruction within the classroom might be improved.

Holding Schools Accountable

In the 1970s complaints apparently reached the ears of state officials and legislators, who began to seek ways to better inform themselves about the state of education and to hold the schools accountable. State after state enacted legislation requiring schools to administer new achievement tests, usually for two reasons: (a) to determine the level of basic skills at various grade levels, and (b) to provide a basis for remediation in schools where it is needed. The "back to basics" movement no doubt helped persuade legislators to vote for the passage of these bills.

State after state enacted legislation requiring schools to administer new achievement tests, usually for two reasons: (a) to determine the level of basic skills at various grade levels, and (b) to provide a basis for remediation in schools where it is needed.

The tests that resulted made use of the multiple-choice format and stressed the mastery of basic skills. These tests were generally referred to as Minimum Competency Tests (MCTs). But there was no agreed-upon definition of minimum competency. The tests were intended to assess whatever basic skills the educators and state authorities decided were the "minimum acceptable outcomes of an education" (Winfield, 1987, p. i).

It was not uncommon, in investigating the influences of MCT programs, to find that school officials improved scores by such devices as excluding low-ability students from taking the test (Murnane, 1987). Fortunately, data from the National Assessment of Educational Progress (NAEP) were free from such bias, as are the other studies reported below.

Some Evidence Regarding the Influence of MCTs on Teaching and Learning

A Texas study. A study in Texas (Mangino & Babcock, 1986) investigated the influence of MCTs separately for

basic and high-level skills, using data from the Austin Independent School District in Texas. In 1979, legislation mandated the development and use of MCTs to assess basic skills in reading, writing and mathematics. Mathematics was chosen as the area to investigate because it was thought the most susceptible to changes in instruction.

. . .before the first test administration, supporting materials such as sample items and practice tests were developed and distributed to teachers to help students master the TABS objectives, and the superintendent of schools gave speeches urging the staff to do better in preparing students for TABS every year.

The MCT was called the Texas Assessment of Basic Skills (TABS). It was administered to all public school students throughout the state, beginning in 1979 and 1980. But even before the first test administration, supporting materials such as sample items and practice tests were developed and distributed to teachers to help students master the TABS objectives, and the superintendent of schools gave speeches urging the staff to do better in preparing students for TABS every year. Two dependent measures were obtained, one of basic math skills, and one of high-level skills. Scores on these subtests were used to investigate the effects of the TABS test on efforts to teach the corresponding basic skills. Test data were obtained from the students at the relevant grade levels in 1979, 1981, and 1983. The numbers of students tested was large: 1,789 ninth-grade students were tested in 1979, 1,483 in 1981, and 1,381 in 1983, all from the Austin Independent School District.

There was significant improvement in basic skills that could be attributed to TABS (the MCT). However, improvement took place only during the interval between 1979 and 1981. There was no significant change in basic skills after 1981.

However, it was found that *low* achievers (those with grades of C or D) did gain significantly in basic skills. They improved significantly in the acquisition of high-level skills, while the high achievers did not. Mangino and Babcock attribute this to "a transfer of knowledge and understanding from the stronger foundation in basic skills to high-level mathematics skills" (p. 13). There seems little doubt that the MCT testing, together with the associated changes in instructional methods, produced some substantial changes in student performance.

There seems little doubt that the MCT testing, together with the associated changes in instructional methods, produced some substantial changes in student performance.

New Jersey study. In some of the high-stakes states, the MCT program was vigorously employed as a device for motivating the teaching and learning of basic skills. In New Jersey, for example, it was made clear that the mandated program required not only that high-school graduation depend on one's MCT score; it also required that results of the testing be reported to the press for publication, that scores be released to all school districts, buildings, and classrooms, and that individual scores be reported to students and

to their parents, as was required by the New Jersey Office of State Educational Assessment (1980). Schools failing to achieve certain standards were to be subjected to review and possible recommendations for remedial action. Teachers were, in effect, required to teach what the test measured.

Schools failing to achieve certain standards were to be subjected to review and possible recommendations for remedial action. Teachers were, in effect, required to teach what the test measured.

It should be no surprise to learn that "A significantly larger percentage of students met the statewide minimum standards in all grades and subjects in 1979-80, more than in the previous two years of Minimum Basic Skills testing (6.9 percent increase in mathematics and 3.4 percent in reading). The MCT was certainly driving instruction in New Jersey.

However, state officials soon realized that the focus on minimum skills was too narrow, and that the MCT should be replaced by a test to assess the higher-level skills needed in order to become "productive members of society." The new test, called the High School Proficiency Test (HSPT), consists of three parts—reading, mathematics and writing. The mathematics test included three- and four-step word problems, pre-algebra, and geometry, instead of simple computations and one-step word problems (Koffler, 1987).

In 1986, only the HSPT was administered; by then, it was required for graduation. The results were far beyond expectations: The percentage of students passing the test for 1984, 1985, and 1986, respectively, were 53.6, 56.0, and 71.8. The HSPT by 1986 had had an important influence on the curriculum.

Several studies using data from the National Assessment of Educational Progress (NAEP) suggest that an overemphasis on minimum competencies might detract from learning the skills associated with higher-order thinking.

NAEP studies. Several studies using data from the National Assessment of Educational Progress (NAEP) suggest that an overemphasis on minimum competencies might detract from learning the skills associated with higher-order thinking. In 1982 NAEP reported that performance on items measuring basic skills was not declining, but there was a decrease in performance on items that required more complex cognitive skills. For example, in mathematics, 90 percent of the 17-year-olds could handle simple arithmetic problems (subtraction and addition), but their performance on problems that required understanding of mathematical principles dropped from 62 percent to 58 percent; on more advanced mathematical problem-solving, the drop was from 33 percent to 29 percent.

A more recent NAEP report on mathematics (Dossey, Mullis, Lindquist, & Chambers, 1988) states that "While average performance has improved since 1973, the gains have been confined primarily to lower-order skills . . . . Most students, even at 17, do not possess the breadth and

depth of mathematical proficiency needed for advanced study in secondary school mathematics" (p. 10). And the report on science achievement (Mullis & Jenkins, 1988) states that "At age 17, students' science achievement remains well below that of 1969 . . . . Only 7 percent of the nation's 17-year-olds have the requisite knowledge and skills thought to be needed to perform well in college-level science courses" (p. 6). A slight improvement was noted from 1982 to 1986, but "It must be recognized, . . . that improvements in average performance. . . were largely the result of students' increased knowledge *about* science rather than increased skills in scientific reasoning, which suggests that current reforms tend to be aimed primarily at symptoms rather than the disease" (p. 11).

. . .that improvements in average performance. . . were largely the result of students' increased knowledge about science rather than increased skills in scientific reasoning, which suggests that current reforms tend to be aimed primarily at symptoms rather than the disease" (p. 11).

A Nationwide Study of the Influence of State–Mandated Testing Practices

Purpose of the Study

The primary purpose of this study was to investigate the influence of state-mandated minimum-competency tests (MCTs) on the quality of teaching and learning as it is reflected in NAEP assessments. One hypothesis was that the use of MCTs would result in improvements in the learning of basic skills—elementary abilities necessary for the lower levels of performance in such areas as reading, writing, and arithmetic. A second hypothesis was that any improvement in basic skills would be gained at the expense of higher-order skills—those needed for a deeper understanding of literature, science, math, etc. An emphasis on teaching basic skills, we thought, might interfere with any efforts to increase the development of the higher-order skills.

Plan of the Study

The plan is quite simple, though it involved nationwide samples of students. We planned to compare the performance of students who participated in the 1978 NAEP mathematics assessment with the performance of those who participated in the 1986 NAEP mathematics assessment, using the same set of items as was used on both occasions, for 9-, 13-, and 17-year–old students. The reason for the comparison is that the 1978 assessment occurred *before* the minimum competency tests were in use, and the 1986 assessment occurred *after* the MCTs had been widely used—a before-and-after

We planned to compare the performance of students who participated in the 1978 NAEP mathematics assessment with the performance of those who participated in the 1986 NAEP mathematics assessment, using the same set of items as was used on both occasions. . .

design. Different samples of students were involved in the two assessments, but there is little doubt that the students were equally representative of the nation's schools. In order to make the 1986 data match the 1978 data more accurately, it was decided to select our samples of students entirely on the basis of age, so that all the 9-year-olds were *really* 9, not high 8s or low 10s, for example.

We had determined that the 1978 NAEP assessment could not have been influenced by the MCTs; it occurred too early. We also found that a good many of the states had made use of MCTs for at least two years before the 1986 assessment; thus, the before-and-after design was appropriate for all three age groups.

NAEP Policies in Relation to this Study

Age groups and subject matter. NAEP's policy has been to use three age groups in its assessments—9, 13, and 17—and it has commonly assessed three abilities: reading, writing, and mathematics. We decided to use all three age groups, but we chose mathematics as the subject area because math items tend to be more objective and therefore easier to grade reliably than tests in other areas. Only public-school students served as subjects.

. . .we chose mathematics as the subject area because math items tend to be more objective and therefore easier to grade reliably than tests in other areas.

NAEP provides a link to the past. NAEP makes changes in its test items in order to be up-to-date with respect to what is being taught; however, it also uses certain sets of items repeatedly, in order better to detect gains or losses of students' knowledge and skills. This made it possible for us to use the items that had remained the same from 1978 to 1986. The number of items used were 34 for age 9, 53 for age 13, and 56 for age 17 (see Table 1).

Properties of NAEP items. As students progress in their education, they gradually move from the acquisition of basic skills to what are often referred to as higher-order skills. It is therefore necessary for teachers to make corresponding changes in their teaching. It was also necessary for NAEP to make corresponding changes for all three age groups. We therefore thought it necessary to sort the NAEP items into two categories: those that require only basic skills, and those that require at least some degree of higher-order thinking skills. We asked several competent cognitive psychologists to read the NAEP items and to sort them into the two categories. They found it very difficult to make such distinctions, even within age groups. It

Mathematical skills that had been well-practiced and that could be used automatically would surely be considered routine. However, a few items did require more thinking—enough, we thought, to be called nonroutine.

appeared that there were very few items that might be called higher-order skills, especially for the younger students.

It was decided that the terms "routine" and "nonroutine" would better fit the two categories. Mathematical skills that had been well-practiced and that could be used automatically would surely be considered routine. However, a few items did require more thinking—enough, we thought, to be called nonroutine. We decided to use the terms *routine* and *nonroutine* instead of *basic* and *higher-order*.

Table 1 shows the number of items that were used in both the 1978 and 1986 assessments for the three age groups, and the number of items judged to be routine and nonroutine for the three age groups. Obviously, the nonroutine items are scarce, especially for the 9- and 13-year-olds.

Table 1

Number of NAEP Items and Their Classification

	Total Number of Items	Routine Items	Nonroutine Items
Age 9	34	26	8
Age 13	53	43	10
Age 17	56	34	22

High- and low-stakes states. There is another very important factor in the study, one that has to do with how the states managed their MCTs.

The classification depended on the levels and kinds of influence that were exerted by school officials and teachers by using their MCTs and other devices.

By 1984, many of the 48 states had developed MCT programs of one kind or another (Goertz, 1986, 1988; Winfield, 1987). (Alaska and Hawaii were not included in the study.) Goertz and her colleagues helped us sort the states into three categories with regard to their use of MCTs: high-stakes states, moderate states, and low-stakes states. The classification depended on the levels and kinds of influence that were exerted by school officials and teachers by using their MCTs and other devices.

The first category—the high-stakes states—was composed of those states that, we judged, had not only mandated the use of MCTs; they also required school officials and teachers to set standards in terms of MCT scores for granting diplomas or for promoting students to the next grade—important aspects of education for young students and their parents.

The so-called "moderate stakes" were those judged to be neither high- nor low-stakes states. They professed to use the

MCTs for such purposes as monitoring student performance, remediation of simple faults, or coaching those students who badly needed assistance. However, the primary reason for identifying the moderate states was to make clearer the contrast between the high- and low-stakes states.

The remaining states were judged to be "low-stakes" states. None mandated the use of MCT scores for any specific purpose. Some of the states allowed local options regarding the use of MCTs by county, by district, or by individual school. Three states had no MCTs, and two states (Alaska and Hawaii) were not assessed by NAEP. The number of states in each category is shown in Table 2; the students were, of course, different with regard to their skills and knowledge.

Table 2
Number of Participating States in Each Stakes Category

	High-Stakes	Moderate	Low-Stakes
Age 9	9	6	10
Age 13	9	6	10
Age 17	10	7	11

NOTE: These states were included in the national sample in both the 1978 and 1986 assessments.

Airasian (1988) discussed the conditions under which various outcomes of MCT testing might occur. "Measurement-driven instruction" was defined by Airasian in terms of using high-stakes achievement tests to control instruction, with the assumption that "the higher the stakes the greater the impact on instruction" (p. 6). However, he pointed out that the influence of MCTs could vary widely, depending on the conditions under which the tests were to be used. For example, if standards for passing the test are low, the impact will be low, and the maximum effect will occur when both the stakes and the standards are high. Airasian reported that most tests used for certification fit the "high-stakes low-standard" cell and "the greatest impact on instruction will occur when high standards and high stakes are present." Unfortunately, we had no way of knowing how the states set their standards.

The Analysis of the NAEP Data

As is shown in Table 3, the number of subjects was in the thousands for each of the age groups. Obviously, there is no lack of data.

Table 3
Number of Students Tested in Each Stakes Category

	1978			1986		
	High Stakes	Moderate	Low Stakes	High Stakes	Moderate	Low Stakes
Age 9	4053	4084	2782	1751	1388	1825
Age 13	6219	6897	4793	1707	1538	1925
Age 17	7078	7013	5194	1253	1001	1092

NOTE: These students represent those from public schools of the correct age who answered at least one of the items in a state that participated in either 1978 or 1986.

In order to test our hypotheses about the influences of MCTs on teaching and learning, we made use of data that had been obtained by NAEP at the assessments in 1978 and 1986. These dates were chosen because the 1978 assessment preceded any MCT testing, and the 1986 assessments followed a period of two or more years when the MCTs were widely used. The basic data consist of the percents of students who answered items correctly.

The analysis of the NAEP data involved 18 different combinations of students and items; each combination included (1) students from one of the three age groups, (2) students from high-stakes, moderate, or low-stakes states, and (3) a set of routine or nonroutine items (3 x 3 x 2 = 18). For each of these sets, we calculated the average percent of students who passed each test item.

An example of the procedure is shown in Table 4. The data are from *9-year-old students*, who come from *high-stakes states*, and the items are *nonroutine*.

The first column in Table 4 is merely a listing of the numbers for eight NAEP items. They are listed in the order of percents correct. The second column presents the percent of correct answers for each of the math items by students at the 1978 assessment, and the third column contains the percent of correct answers to the same items at the 1986 assessment.

The fourth column contains for each item the difference between the two percents (the 1986 percent minus the 1978 percent). Below the last column is the average of the eight differences. For this set of items, the answer is 6.12. This number indicates the extent to which 9-year-olds from high-stakes states, who took the nonroutine (more difficult) items, showed improvement in math performance compared with similar students who took the same items at the 1978 assessments.

Table 4
Percent of 9-Year-Old Public School Students Who Correctly Answered Nonroutine Math Items in High-Stakes States

Item Number	Percent Correct 1978	Percent Correct 1986	Differences 1986-1978
1	22.14	18.23	-3.91
2	25.16	35.26	10.10
3	27.89	52.17	24.28
4	33.69	31.77	-1.92
5	35.93	37.86	1.93
6	49.04	60.65	11.61
7	63.24	68.45	5.21
8	69.58	71.25	1.67
		Average Difference	6.12

The problem solved (Table 4) is one of 18 similar problems. Its answer (6.12) is at the top of the nonroutine-items column in Table 5.

Table 5
Influences of State–Mandated Testing in Public Schools: The Average Differences in Number of Students Correctly Answering Mathematics Items, from 1978 to 1986

	ROUTINE ITEMS	NONROUTINE ITEMS
	9-Year-Olds	
High–Stakes States	4.44*	6.12*
(Moderate States)	(2.00*)	(2.88*)
Low-Stakes States	-3.46*	1.58*
High-Low Differences	7.90	4.54
	13-Year-Olds	
High-Stakes States	3.27*	-1.34*
(Moderate States)	(6.49*)	(1.26*)
Low-Stakes States	0.17	-5.98*
High-Low Differences	3.10	4.64
	17-Year-Olds	
High-Stakes States	1.92*	0.68*
(Moderate States)	(1.25*)	(-0.24)
Low-Stakes States	1.30*	-1.22*
High-Low Differences	0.62	1.90

* 1986-1978 difference is significant at the 95 percent confidence level.

Some Interpretations of Our Findings

Table 5 presents the routine and nonroutine items separately because the two types of items assess different abilities: Routine items were intended to assess basic skills, and the nonroutine items were to assess higher-order skills. It turned out that there were relatively few nonroutine items (see Table 1). Any differences may be difficult to account for.

Table 5 also separately presents another set of results: those from the high-stakes and low-stakes states for 9-, 13-, and 17-year-old students. The high-stakes states were the states that mandated the use of MCTs. The school officials were those who set standards for granting diplomas and

promoting students. The high stakes are no doubt more important in improving classroom teaching than are the routine and nonroutine items alone.

The high stakes are no doubt more important in improving classroom teaching than are the routine and nonroutine items alone.

It should be noted that the presence of an asterisk indicates that the difference in average percent correct between 1986 and 1978 is statistically significant at the 95-percent confidence level. There are two instances where the asterisk is missing in Table 5: when the 13-year-olds were in low-stakes states, and when the 17-year-olds were in moderate states.

The Routine Items. The numbers in the routine items column of Table 5 were obtained in the way that was demonstrated in Table 4. They are the differences between the 1978 and the 1986 assessments, which are the average percents-correct statistics. The same items were used in both years, but the 1986 students performed better, perhaps because of the influence of those who made use of the MCTs and the high-stakes states.

As the top cell in the routine items (Table 5) column shows, the average percent correct (4.44) was significantly greater for students in the high-stakes states than for those in the low-stakes states (-3.46). The high-low difference is high (7.90). This may indicate that the high-stakes states did something to improve the teaching and learning of basic skills, at least for the 9-year-olds.

It is apparent that the high-low differences are smaller for the 13-year-olds, and still smaller for the 17-year-old students: 3.10 for 13-year-olds and 0.62 for the 17-year-old students. Obviously, the differences were smaller as the 1978 students had more (and perhaps better) instruction.

The Nonroutine Items. The top of the second column of Table 5 shows that the effects of the high-stakes states must have resulted in a still-better performance on the part of the 1978 9-year-olds: The change from 1978 to 1986 for the average percent correct was 6.12, which is high, compared with the 4.44 for the routine items.

The use of MCTs and state-mandated requirements are, of course not the only influences on student performance. Teachers no doubt differ in their teaching styles, and most parents help their children in one way or another.

The pattern for the nonroutine column is somewhat like that of the routine items, except for a negative high-stakes difference (-1.34) for the 13-year-olds and a somewhat higher one (0.68) for the 17-year-olds. The small number of nonroutine items may account for the wide variations in the 1978-1986 differences.

The use of MCTs and state-mandated requirements are, of course not the only influences on student performance. Teachers no doubt differ in their teaching styles, and most parents help their children in one way or another. Some teachers may merely teach the conventional mathematical procedures, while others teach for mathematical

reasoning and understanding. Elementary school teachers and high school teachers necessarily teach in quite different ways. Whatever the causes, the 1986 younger students apparently had profited more from the MCTs and the high-stakes states than the older students did.

It seems reasonable to conclude that the use of MCTs can have desirable influences on the performance of young students as measured by NAEP—especially when high-stakes conditions prevail. It also seems reasonable to assume that for teen-age students too much emphasis on teaching basic skills may indeed interfere with the teaching of higher-order thinking skills. If teaching basic skills interferes with acquiring nonroutine skills, they would surely interfere with teaching more advanced thinking abilities.

APPENDIX: Some Broader NAEP Methods And Interests

NAEP makes use of a variety of procedures for obtaining and analyzing data. The conventional method of test-taking is based on the number of items answered correctly by each student. However, each student participating in the study may have received only one or two of the items in the pool of routine and nonroutine items that was selected. Thus, although the total sample of students in each grade was in the thousands, each item was received by several hundred students. For each item, the percentage of students who answered the item correctly was found (see columns 2 and 3 in Table 4). The near-percent-correct was then calculated for each set of routine and nonroutine items, as was the mean number of students who answered each item. These numbers are the P+ and N seen in tables 6 to 9.

. . .although the total sample of students in each grade was in the thousands, each item was received by several hundred students.

Asterisks and Other Symbols

Asterisks are placed at the ends of certain rows in tables to indicate that the 1978-1986 differences are significant at the 95-percent-perfect confidence level (see the footnote to Table 6).

There are other symbols, the purpose of which is to identify pairs (or subgroups) within groups that are significant at the 95-percent-confidence level (see the second subscript below Table 6). The first symbol used (other than an asterisk) is named "double dagger," and two other symbols to be used are called *omega* and *florin*. Such symbols are needed to identify significant pairs in race demography.

Some Demographic Studies

As was made clear in Table 5, there were two important influences in the development of teaching and learning. Most important to this paper was the introduction of MCTs (Minimum Competency Tests) and the high-stakes states. School officials in many states attempted seriously to set higher standards for granting diplomas and/or promoting students to the next grade, while the low-stakes states were those where little or no effort was made by school officials to see that students' work was well-done.

The other influence resulted from an effort to sort the NAEP math items into two categories: (1) those that required only basic skills, and (2) those that required at least some degree of higher-order-thinking abilities.

Reviewers have suggested that making use of demographic groups would be useful to broaden understanding of the procedures. Two demographic groups will be used in the procedures: *Gender* and *Race* (Ethnicity). The Gender group can of course only divide into two subgroups: Male and Female. The Race group will be composed of Whites, Blacks, and Hispanics.

Reviewers have suggested that making use of demographic groups would be useful to broaden understanding of the procedures.

Comments on Tables 6 to 9

Each of the following tables describes various influences on the performance of students who are being assessed by NAEP. The situations vary quite widely, with the high-stakes and the low-stakes states, the routine and the nonroutine items, the age of the students, and the gender and race of the students. But the items are the same for the 1978 and the 1986 NAEP assessments, and for each of the three age groups.

Tables 6, 7, 8, and 9 were prepared for use in controlling some of these factors, especially the high- and low-stakes, and the routine and nonroutine items:

Table 6. High-stakes states and routine items
Table 7. Low-stakes states and routine items
Table 8. High-stakes states and nonroutine items
Table 9. Low-stakes states and nonroutine items

Comments on Table 6: High-Stakes States and Routine Items

This table presents the results of both the high-stakes states and the routine items. The 1978 column presents the percents that are typical of NAEP math assessments, and the 1986 column shows the percents of the items that were influenced by the high-stakes states and the use of the routine items (those that had been reasonably well-practiced by the students). As shown by the asterisks, the 1986 and 1978 differences are all significant at the 95-percent level. The phrase "by subgroup" in the title means that the findings can be reported separately for demographic subgroups, e.g., male and female students, or for students who are White, Black, or Hispanic.

Table 6

Changes from 1978 to 1986 in the Percent of Routine Items Answered Correctly in High-Stakes States by Subgroup

		1978		1986		1986-1978
Age	**Subgroup**	**P+**	**N**	**P+**	**N**	**Difference**
			GENDER			
9	Male	53.41‡	340	57.54‡	290	4.13*
	Female	55.78‡	339	60.62‡	293	4.84*
13	Male	63.66	311	66.58	291	2.92*
	Female	63.71	303	67.37	275	3.66*
17	Male	68.14	306	70.24	288	2.10*
	Female	67.71	331	69.47	346	1.76*
			RACE/ETHNICITY			
9	White	59.34‡'	417	63.92‡'	360	4.58*
	Black	44.63‡	199	46.47‡ ƒ	99	1.84
	Hispanic	45.90 '	56	51.23 'ƒ	90	5.33
13	White	68.59‡'	410	70.45‡'	284	1.86*
	Black	48.70‡	152	61.09‡ ƒ	169	12.39*
	Hispanic	50.39 '	48	57.60 'ƒ	99	7.21*
17	White	73.52‡'	450	74.73‡'	410	1.21*
	Black	49.16‡	143	54.52‡ ƒ	153	5.36*
	Hispanic	50.81 '	36	59.05 'ƒ	60	8.24

* -- 1986-1978 difference is significant at the 95 percent confidence level.

‡ ' ƒ The difference between the 2 subgroups with the same mark (within a category at each age) is significant at the 95 percent confidence level.

The 1986-1978 difference column also shows that the amount of change varies: The differences become smaller as the age of the student increases. The differences for the 9-year-olds were 4.13 for the males and 4.84 for the females; but for the 17-year-olds the numbers were 2.10 for males and 1.76 for females. The differences between males and females are not great, but the smaller difference apparently goes with higher age. Table 5 shows a similar phenomenon. The only significant differences between the subgroups involve 9-year-old males and females; the reason is not obvious.

. . .the percents in the 1986 column are all higher than those in the first column. The causes may well be that the high-stakes states had provided incentives for better teaching and learning, and the routine items had provided opportunities for better solving of mathematics items.

Gender. An interesting aspect of this part of the table is that the percents in the 1986 column are all higher than those in the first column. The causes may well be that the high-stakes states had provided incentives for better teaching and learning, and the routine items had provided opportunities for better solving of mathematics items. It is also of interest that 1986-1978 differences are much smaller for the 17-year-olds than for the 9-year-olds—a common but not understood phenomenon.

The only significant difference between the two subgroups involves 9-year-old males and females.

Race/Ethnicity. This group is composed of three subgroups: White, Black, and Hispanic students in three age groups. It is obvious from the title that the 1986 students had come from the high-stakes states, and that a considerable proportion had answered correctly the routine items—the items that had been frequently practiced. But there had been considerable differences among the three subgroups, even those with different age groups. For example, Table 6 shows that all of the White student percents-correct are substantially higher than those for the Black and Hispanic students, and the age groups differ widely.

It is obvious from the title that the 1986 students had come from the high-stakes states, and that a considerable proportion had answered correctly the routine items—the items that had been frequently practiced.

There are differences that must have been made by the students concerned with the 1986 increases in P+. For example, the 13-year-old Blacks moved from 48.70 in 1978 to 61.09 in 1986; the difference is 12.39. The 13-year-old Hispanics moved from 50.39 to 57.60—a difference of 7.21. And the 13–year–old Whites at the same time moved from 68.59 to 70.49—a difference of 1.86.

Similar patterns can be found for the 9- and 17-year-olds: Whites are well above both Blacks and Hispanics. When comparing only the Blacks and Hispanics, the latter are usually the highest.

If we look at the 1986–1978 Differences, we find in the 9-year subgroup that both Whites and Hispanics made substantial gains—4.58 for Whites and 5.33 for Hispanics. In the

13-year subgroup, we find even larger gains: 12.39 for Blacks and 7.21 for Hispanics. And in the 17-year subgroup we find that the largest gains are 8.24 for Hispanics and 5.36 for Blacks. These differences are large but not significant.

Analyses of comparisons such as this can best be carried out by working with pairs of numbers. The difference between two such subgroups with the same mark (within a category at each age) is significant at the 95-percent confidence level. The symbols used should be double dagger, omega, and caret.

Analyses of comparisons such as this can best be carried out by working with pairs of numbers. The difference between two such subgroups with the same mark (within a category at each age) is significant at the 95-percent confidence level.

Comments on Table 7: Low Stakes and Routine Items

The 1986-1978 Difference in Table 7 looks like the reverse of the numbers in Table 6, where the Difference numbers go from high (4.13 and 4.84 for 9-year-olds) to low (2.10 and 1.76 for 17-year-olds). By contrast, the Table 7 1986-1978 Difference numbers go from low (-2.66 and 4.31 for 9-year-olds) to high (2.29 and 0.48 for 17-year-olds).

The reasons are clear. Students in the 1986 assessment were *not* in the high-stakes states; they were in *low*-stakes states, where there were no MCTs, and no officials and teachers to set standards in the form of MCT scores to grant diplomas or to promote students to the next grade. The routine items apparently cannot compensate for the loss of high-stakes states.

Both males and females shared in the reversal of the numbers in Table 7, but there is a small difference between males and females with regard to the size of the 1986-1978 differences.

Gender. Both males and females shared in the reversal of the numbers in Table 7, but there is a small difference between males and females with regard to the size of the 1986-1978 differences. The females had fewer correct answers. The differences were -4.31 for females and -2.66 for males at the 9-year age. There were fewer differences with -0.86 for females and 1.05 for males at the 13-year age. And the 17-year-olds were apparently able to disregard certain aspects of the situation, and to turn to the implications of positive interpretations: differences of 2.29 for males and 0.48 for females. However, these differences are not significant.

Race (Ethnicity). Table 7 displays White, Black, and Hispanic subgroups. The White students are far ahead of the Black and Hispanic students with regard to percents of routine items; for the 9-year-old White-Black pair the differences are 20.19 (in 1978) and 11.73 (1986). And for the White-Hispanic pair the differences are 14.91 (in 1978) and 11.21 (in 1986). There is still plenty of room for the Black and Hispanic students to overtake the Whites.

Table 7

Changes from 1978 to 1986 in the Percent of Routine Items Answered Correctly in Low-Stakes States by Subgroup

		1978		1986		1986-1978
Age	**Subgroup**	**P+**	**N**	**P+**	**N**	**Difference**
			GENDER			
9	Male	59.20	229	56.54	318	-2.66*
	Female	60.31	227	56.00	287	-4.31*
13	Male	65.05	241	66.10	322	1.05
	Female	66.34	230	65.48	326	-0.86
17	Male	72.41‡	226	74.70‡	269	2.29*
	Female	69.80‡	237	70.28‡	286	0.48
			RACE/ETHNICITY			
9	White	61.94‡‡	371	59.37‡‡	396	-2.57*
	Black	41.75‡	65	47.64‡	64	5.89
	Hispanic	47.03 ˚	16	48.16 ˚	94	1.13
13	White	68.27‡‡	390	68.43‡‡	391	0.16
	Black	50.80‡ ƒ	57	59.39‡ ƒ	152	8.59*
	Hispanic	53.90 ˚ƒ	20	55.87 ˚ƒ	88	1.97
17	White	73.65‡‡	409	75.94‡‡	398	2.29*
	Black	45.87‡	35	56.71‡	59	10.84*
	Hispanic	53.78 ˚	13	59.10 ˚	75	5.32

* -- 1986-1978 difference is significant at the 95 percent confidence level.

‡ ˚ ƒ The difference between the 2 subgroups with the same mark (within a category at each age) is significant at the 95 percent confidence level.

If we compare these numbers with low-stakes states (Table 7) there is a very poor match: for 9-year-olds, the 1986-1978 differences are -2.57, 5.89, and 1.13, and for 17-year-olds they are 2.29, 10.84, and 5.32.

A pair-by-pair inspection of Table 7 shows only one instance where the 1986 number was lower than the corresponding 1978 number: The 1978 number is 61.94 and the 1986 number is 59.37—a difference of -2.57. In all other instances the 1986 numbers were higher.

The 1986-1978 differences column shows clearly the instances where the 1986 numbers are substantially higher. The largest such number was 10.84, for 17-year-old Blacks; next was 8.59, for 13-year-old Blacks; and next was 5.89 for 9-year-old Blacks. It appears that the Blacks are more inclined to improve their status in mathematics than the Hispanics.

One must guess as to the reasons for these changes, but it is obvious that White students in 1978 were far ahead of Black and Hispanic students. At the 9-year-old level, the Whites' percentage of the routine items was 61.94, while the Blacks and Hispanics were at 41.75 and 47.03. In 1986, the percents answered correctly were 59.37 for Whites (down by -2.57) and 47.64 and 48.16 for the Black and Hispanics—increases of 5.89 and 1.13.

At the 17-year age, it was clear that the Black students were ahead of both White and Hispanic students in the 1986-1978 Difference column. The Whites had gained 2.29, the Hispanics 5.32, and Blacks had gained 10.84.

An analysis of comparisons such as this can best be carried out by working with pairs of numbers. The difference between two such subgroups with the same mark (such as the double dagger) is significant at the 95-percent confidence level. The symbols most likely to be used are the double dagger, omega, and caret.

Comments on Table 8: High-Stakes States and Nonroutine Items

One might expect that the influence of the high-stakes states would overpower any effects that are produced by the few nonroutine items. Such is not the case.

One might expect that the influence of the high-stakes states would overpower any effects that are produced by the few nonroutine items. Such is not the case. The 1986-1978 difference column seems to have little to do with the high-stakes states.

Gender. Including all but one of the differences from age 9 to age 17, the differences range from -1.64 to 1.16—a range of 2.80, which is not great for 9- to 17-year-old students. But there is one remaining large difference:

Table 8

Changes from 1978 to 1986 in the Percent of Nonroutine Items Answered Correctly in High-Stakes States by Subgroup

		1978		1986		1986-1978
Age	**Subgroup**	**P+**	**N**	**P+**	**N**	**Difference**
			GENDER			
9	Male	42.17‡	337	43.33‡	288	1.16
	Female	39.46‡	338	50.42‡	299	10.96*
13	Male	48.29‡	321	47.12	289	-1.17*
	Female	49.44‡	308	47.80	277	-1.64*
17	Male	54.67‡	233	53.80‡	268	-0.87
	Female	48.91‡	238	47.50‡	284	-1.41*
			RACE/ETHNICITY			
9	White	45.56‡"	416	51.55‡"	363	5.99*
	Black	32.65‡ ƒ	196	35.63‡	99	2.98*
	Hispanic	23.59 "ƒ	58	38.00 "	93	14.41*
13	White	53.27‡"	411	51.28‡"	284	-1.99*
	Black	35.92‡	164	41.38‡ ƒ	171	5.46*
	Hispanic	34.79 "	49	37.27 "ƒ	98	2.48
17	White	53.36‡	415	53.85‡"	395	0.49
	Black	36.41‡	37	36.51‡	58	0.10
	Hispanic	41.69	13	38.97 "	75	-2.72

* -- 1986-1978 difference is significant at the 95 percent confidence level.

‡ " ƒ The difference between the 2 subgroups with the same mark (within a category at each age) is significant at the 95 percent confidence level.

10.96. This is well out of the usual range. And it was produced by 9-year-old females. How does one account for this peculiar set of differences? One might expect that the influence of high-stakes states would overpower whatever negative effects might be produced. The nonroutine items not only require some degree of thinking, but they are few in number.

Along with the high-stakes states, the nonroutine items were included in the definition of Table 8. Table 1 indicates that the age-9 students must deal with 34 items, 26 of which are routine and eight are nonroutine. The *routine* items are those that had been well-practiced and could be solved more or less automatically. The *nonroutine* items are those that require higher-order thinking skills. One might expect that the influence of high-stakes states would overpower any negative effects that are produced by the nonroutine items. Such is apparently not the case.

Race (Ethnicity). The percent of nonroutine items answered correctly in high-stakes states is unusually low in comparison with the preceding tables. For example, here is a sample of Table 6.

Sample of Table 6

	1978	1986	1986-1978 Difference
White	68.59	70.45	1.89
13 Black	48.70	61.09	12.39
Hispanic	50.39	57.60	7.21

Compared with Table 8

White	53.27	51.28	-1.99
13 Black	35.92	41.38	5.46
Hispanic	34.79	37.27	2.48

In spite of the nonroutine items and the high-stakes states, the Table 8 numbers are lower than those of Table 6. The differences for the 1978 column are about 14 for the 1978 column and about 20 for the 1986 column. The reason for the difference is probably the nonroutine items; the number of items that require higher-order thinking are no doubt too small.

Comments on Table 9: Low-Stakes States and Nonroutine Items

Table 9 lacks both the high-stakes states and lacks information on any substantial way to improve the performance of students. There seems to be little order to the Table 9 1986-1978 differences, as has been suggested. In the light of the comments about Tables 7 and 8, it seems that Table 9 makes little sense.

Table 9

Changes from 1978 to 1986 in the Percent of Nonroutine Items Answered Correctly in Low-Stakes States by Subgroup

		1978		1986		1986-1978
Age	**Subgroup**	**P+**	**N**	**P+**	**N**	**Difference**
			GENDER			
9	Male	45.69	235	47.74‡	315	2.05*
	Female	45.59	237	46.50‡	288	0.91
13	Male	51.66‡	241	48.10‡	327	-3.56*
	Female	53.18‡	239	44.83‡	323	-8.35*
17	Male	51.22‡	311	50.35‡	283	-0.87
	Female	45.18‡	331	47.50‡	340	2.32*
			RACE/ETHNICITY			
9	White	47.91‡	367	51.96‡"	395	4.05*
	Black	28.28‡	83	31.39‡	62	3.11
	Hispanic	37.84	18	35.94 '	98	-1.90
13	White	54.93‡"	397	49.29‡"	395	-5.64*
	Black	40.05‡	60	38.17‡	152	-1.88
	Hispanic	33.66 '	20	38.10 '	87	4.44
17	White	52.49‡"	466	52.85‡"	406	0.36
	Black	32.59‡	136	35.44‡	150	2.85*
	Hispanic	37.03 '	35	38.77 '	57	1.74

* -- 1986-1978 difference is significant at the 95 percent confidence level.

‡ " ' The difference between the 2 subgroups with the same mark (within a category at each age) is significant at the 95 percent confidence level.

Bibliography

Arasian, Peter W. (1988). Measurement driven instruction: A closer look. *Educational Measurement: Issues and Practics, 7.*

Coley, R. J. & Goertz, M. E. (1990). *Educational standards in the 50 states.* Princeton, NJ: Educational Testing Service.

Donlon, T. F., Hilton, T. L., & Schrader, W. B. (1978). *Designing a test plan for the national longitudinal study "High School and Beyond."* Princeton, NJ: Educational Testing Service.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. M., & Chambers, P. L. (1988). *The mathematical report card: Are we measuring up?* Princeton, NJ: Educational Testing Service.

Goertz, M. E. (1986/1988). *State educational standards: A 50-state survey.* Princeton, NJ: Educational Testing Service.

Hilton, T. L. & Rhett, H. (1973). The base year survey of the national longitudinal study of the high-school class of 1972. *The Final to the U.S. Department of Health, Education, and Welfare, and National Center for Educational Statistics.* Princeton, NJ: Educational Testing Service, 1973.

Koffler, S. L. (1987). Assessing the impact of a state's decision to move from minimum competence testing toward higher-level testing for graduation, *Education for Evaluation and Policy Analysis, 9,* 325-336.

Lindquist, E. F. (1954). *The Iowa electronic test processing.* Paper presented at the Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Mangino, E., & Babcock, M. A. (1986). *Minimum competence testing: Helpful or harmful high level skills.* Paper presented at the Annual Meeting of American Educational Research Association, San Francisco, CA.

Mullis, I., and Jenkins, L. B. (1988). *Science Learning Matters: An Overview of the Science Report (NAEP).* Princeton, NJ: Educational Testing Service.

Raisen, S. & Jones, L. (1988). Indicators of precollege education in science and mathematics. *National Academy Press.*

Popham, W. M. (1983). Measurement as an instructional catalyst. In *Directions for Testing and Measurement: Measurement, Technology, and Individuality in Education.* Edited by R. B. Ekstrom. San Francisco, CA: Jossey-Bass.

Rock, D., Ekstrom, R. B., Goertz, M. E., Hilton, T. L. and Pollock, M. B. (1985). *Factors associated with decline of test scores of high school seniors, 1972-1980: A study of excellence: Educational policies, school quality, and student outcomes.* Princeton, NJ: Educational Testing Service.

Tyler, R. A. (1934). Constructing achievement tests. *Columbus State University.*

Winfield, L. F. (1987). The relationship between minimum competency testing programs and students' reading ability. *Implications from the 1982-84 national assessment of educational progress of reading and writing.* Princeton, NJ: Educational Testing Service.

Womer, F. B. (1981). State Level testing: Where have we been may not tell us where we are going. In *New directions for testing and measurement: Beyond accountability.* Edited by D. Carlson, Jossey Bass.